OpenAssistant Conversations - Democratizing Large Language Model Alignment

Abstract

Aligning large language models (LLMs) with human preferences has proven to drastically improve usability and has driven rapid adoption as demonstrated by ChatGPT. Alignment techniques such as supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) greatly reduce the required skill and domain knowledge to effectively harness the capabilities of LLMs, increasing their accessibility and utility across various domains. However, state-of-the-art alignment techniques like RLHF rely on high-quality human feedback data, which is expensive to create and often remains proprietary. In an effort to democratize research on large-scale alignment, we release OpenAssistant Conversations, a human-generated, human-annotated assistant-style conversation corpus consisting of 161,443 messages in 35 different languages, annotated with 461,292 quality ratings, resulting in over 10,000 complete and fully annotated conversation trees. The corpus is a product of a worldwide crowd-sourcing effort involving over 13,500 volunteers. Models trained on OpenAssistant Conversations show consistent improvements on standard benchmarks over respective base models. We release our code and data under a fully permissive licence.